The Research and Teaching Corpus of Spoken German – FOLK

Thomas Schmidt
Institut für Deutsche Sprache
R5, 6-13, D-68161 Mannheim
E-mail: thomas.schmidt@ids-mannheim.de

Abstract

FOLK is the "Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)" (eng.: research and teaching corpus of spoken German). The project has set itself the aim of building a corpus of German conversations which a) covers a broad range of interaction types in private, institutional and public settings, b) is sufficiently large and diverse and of sufficient quality to support different qualitative and quantitative research approaches, c) is transcribed, annotated and made accessible according to current technological standards, and d) is available to the scientific community on a sound legal basis and without unnecessary restrictions of usage. This paper gives an overview of the corpus design, the strategies for acquisition of a diverse range of interaction data, and the corpus construction workflow from recording via transcription an annotation to dissemination.

Keywords: speech corpus, conversation analysis, spontaneous speech

1. Introduction

Data of spontaneous verbal interactions are the empirical basis of many types of research in linguistics and speech technology. For the German language, however, only few such data have been made available to the scientific community so far. Those that have been made available are typically either relatively small and old (such as the Freiburger Korpus, Engel & Vogel 1975, and the Korpus Dialogstrukturen, Berens et al. 1976) or restricted to a certain discourse domain (such as the GeWiss corpus of academic speech, Fandrych et al. 2012) or a certain speaker type (such as the KidKo corpus of young speakers with a multi-ethnic background in Berlin, Wiese 2012 and several multilingual corpora provided by the HZSK, Hedeland et al. 2014). Constructing and publishing an up-to-date and broadly diversified corpus of spoken German can therefore be expected to benefit a great variety of users. The project “Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)” (eng.: research and teaching corpus of spoken German) has therefore set itself the aim of building a corpus of German conversations which:

a) covers a broad range of interaction types in private, institutional and public settings,
b) is sufficiently large and diverse and of sufficient quality to support different qualitative and quantitative research approaches,
c) is transcribed, annotated and made accessible according to current technological standards,
d) is available to the scientific community on a sound legal basis and without unnecessary restrictions of usage.

The FOLK project started in 2008. By today, an initial set of data comprising over 100h of recordings and close to 1,000,000 transcribed tokens has been completely processed and published via the Database for Spoken German (DGD2, Schmidt 2014). This paper describes the corpus design, its current composition and strategy for future extension (section 2) as well as the corpus construction workflow (section 3) and the dissemination methods for completed data (section 4).

2. The corpus

2.1 Corpus design

The primary dimension in the design of FOLK is a stratification according to interaction types. We aim at covering a maximally diverse range of verbal communication in private, institutional and public settings.

	# interactions	# tokens	hours
Private interaction			
Coffee table conversation	7	89281	08:01
Couple conversation	3	20980	02:42
Familiy conversation	2	23414	01:50
Conversation among friends	1	24744	02:17
Conversation among students	4	42295	03:07
Conversation on a holiday trip	2	5477	00:29
Conversation during housekeeping	1	5228	00:21
Adults playing parlor games	2	64968	06:42
Playing games with children	4	40514	05:09
Reading to children	6	18901	02:59
Interaction in school / university			
Lesson at a commercial high school	8	51760	07:00
Lesson at a vocational school	7	50050	07:13
Oral exams at a university	19	98592	10:21
Feedback among teachers	1	5991	00:24
Interaction at the workplace			
Meeting in a social institution	3	85256	07:34
Shift change at a hospital	8	28108	02:38
Training in an aid organisation	9	15217	01:36
Conversation at a police station	9	27515	03:12
Public interaction			
Mediation talks	2	102535	10:29
Other interaction types			
Maptasks	25	64257	07:16
Biographic Interviews	14	100569	09:33

Table 1: Interaction types

This includes, for instance, data from educational institutions (classroom discourse, academic exams, etc.), from the workplace (staff meetings, training, etc.), from service encounters, from the private domain (e.g. "coffee-table" conversation, interaction during every-day activities like cooking), and from the public space (e.g. panel discussions).
We also attempt to control for some secondary variables, like regional variation, sex and age of speakers, in order to achieve a corpus out of which balanced samples can be extracted.
Since we are interested in documenting communication practices in their entirety, we always record and transcribe full interactions, rather than selected excerpts.
Initially, most recordings were audio only. Acknowledging that visible forms of communication (gestures, mimics, actions) are often as crucial to interaction as their audible counterparts, however, we are now attempting to make video recordings wherever the field conditions and the legal circumstances allow.
Table 1 gives an overview of the interaction types included in the latest release of the corpus.
With over 100h of recordings and almost 1 million transcribed tokens, this release marks the end of the first corpus construction phase in which priority was given to establishing the corpus creation workflow and quickly building up an initial, diversified body of data. Since we concentrated on easily accessible data in the first phase, the current release is markedly biased towards data from south western Germany (see table 2) and towards younger speakers with a higher education background.

Region	*# interactions*	*# tokens*	*hours*
North Low German	8	49611	04:28:39
Pommeranian	1	1352	00:10:56
Margravian	4	16386	02:01:60
Westphalian	1	1545	00:11:47
Eastphalian	2	6068	00:40:33
Upper Saxon	23	127190	13:08:27
Ripuarian	8	24822	02:30:52
Hessian	8	110359	10:32:28
Thüringen	2	13934	01:25:56
Moselle Franconian	1	8986	00:41:15
Rhine Franconian	43	310023	32:15:57
East Franconian	11	31330	03:38:07
Swabian	5	118119	11:56:12
Alemannic	10	81297	09:38:13
Bavarian	7	54977	06:52:25
not documented	3	9653	00:59:38

Table 2: Dialect regions

In the current corpus construction phase, we are therefore maintaining the priority on diversification with respect to interaction types, but are also paying more attention to balancing the corpus in terms of regional variation and speaker properties.

2.2 Data acquisition

In order to obtain spontaneous speech data of the diversity aimed for in FOLK, it is crucial to have a suitably diverse field access.
We partly reused material recorded in other projects of the institute (most importantly from the corpus "Deutsch heute", Brinckmann et al. 2008), but mostly acquire new material in the project itself.
So far, employing local university students who use their private networks for gaining field access to different communication domains has been the main approach to diversifying field access. In addition, we also solicited data donations from (usually completed) external projects which furnished us with data from specific communication domains (academic discourse from the GeWiss project, Fandrych et al. 2012, police interrogations from a PhD project, Hee 2012) and from a specific area in Germany (Northern German variants from the SiN project, Kellner et al. 2012).
As we are extending the range of interaction types and beginning to systematically address the problem of regional stratification, it becomes more and more important to unlock further data sources outside the project itself. We are currently processing data where the recordings and initial transcriptions were commissioned to external projects with expertise in oral corpora (the HZSK in Hamburg[1] and the KgSR project in Bochum[2]) as a way of "outsourcing" the field access problem and obtaining data from other parts of Germany.
By making the crucial components of the corpus creation workflow – such as metadata and consent forms, the transcription software and guidelines used in the project – available to the scientific community, we hope to encourage more and more external researchers to create data (also) usable for FOLK in the future.

3. Corpus creation workflow

3.1 Preparation of recordings, anonymization

Recordings are typically made with field recorders as the maximally simple and minimally obstructive technical setup. Where the field conditions allow or the interaction type makes it necessary, more sophisticated recording setups are used – for instance, classroom discourse is usually recorded on video in order to facilitate speaker assignment of utterances; for recordings of driving lessons (not yet included in the published version), we used an additional camera to capture the participant's view out of the car's front shield.
Recording assistants are instructed to gather metadata about the interactions and speakers as well as signed consent forms immediately before the actual recording takes place. The metadata forms capture salient characteristics of the speech event, of the recording conditions, and of the participating speakers. The consent

[1] see http://www.corpora.uni-hamburg.de/
[2] see http://www.ruhr-uni-bochum.de/kgsr/

forms authorize masked versions of the recordings and transcripts (see below) to be published via the internet and used for research and teaching purposes. Only data that are complete in this respect are considered for further processing in the project.
Recordings from the field are optimized (i.e. normalized, denoised) by a trained technician in the project, taking care not to tamper with the authenticity of the recording (e.g. not overly reducing background noise), and finally converted to the standards defined by the Archive for Spoken German in which the project is located (meaning, among other things, that audio is stored in 48kHz WAV files).
An anonymization template is then created which marks all locations in the recording where a person or place name or other information occur which would allow a direct identification of the speakers involved. These locations are then replaced by a brown noise in the recordings. The template also contains a pseudonymization table that is later used during transcription to consistently replace the person and place names with suitable pseudonyms.

3.2 Transcription

For transcription, we follow the conventions of the GAT system (Selting et al. 2009) which uses a modified orthography ("literarische Umschrift" – literal transcription) to represent common phenomena of spontaneous speech (such as elisions, contractions, etc.) and pronunciations deviating from the standard (such as dialectal forms). Non-verbal articulations (laughing, coughing, audible breathing etc.) and actions (e.g. writing on a blackboard) are also noted as long as they are alternative, rather than simultaneous, to speech.

0001	**VK**	°h dann is alles was dann anschließends (.) da raus folgt °h öh
0002		(0.65)
0003	VK	schlichtweg (.) nich brauchbar
0004		(0.38)
0005	**MH**	ne[in]
0006	**VK**	[das] ist das ist die problematik [°h also mein vor]schlag mein vorschlag is
0007	**MH**	[herr kefer herr ke]
0008		(0.45)
0009	**VK**	wir (.) wir lassen des jetz einfach nach dem was ich jetzt gr (.) grade ausgeführt hab wir lassen_s jetzt einfach mal so stehen °h weil ne weitere diskussion können wer nicht führen [wir brauchen den richt]igen wert

Figure 1: Transcript example

Special attention is paid to an accurate measurement of silent pauses and to a precise temporal marking of overlapping speech.
Figure 1 shows an excerpt of a GAT transcript in which most of these phenomena occur.
The choice of this transcription system is motivated, first, by the fact that it is one of the most commonly used in German conversation analysis and related fields and, second, that it requires – at least in its "minimal" version – relatively few interpretative decisions on the transcribers' side, which we aim to avoid both for reasons of efficiency and in order to make the corpus data usable for a wide range of research approaches.
Transcription is usually carried out by student assistants in the project, sometimes also by externally commissioned transcribers (see above), using the software FOLKER (Schmidt 2012). The software takes care of the alignment between transcription texts and media files during the transcription process. It stores transcriptions in an XML format whose underlying data model is largely compatible with other widely used annotation software like EXMARaLDA, ELAN or Praat. Figure 2 shows the XML encoding of the first speaker contribution of the above excerpt.

```
<contribution speaker-reference="VK">
  <time time="0.0"/>
  <breathe type="in" length="1" id="b1"/>
  <w pos="ADV" lemma="dann">dann</w>
  <w n="ist" pos="VAFIN" lemma="sein">is</w>
  <w pos="PIS" lemma="alle">alles</w>
  <w pos="PIS" lemma="was">was</w>
  <w pos="ADV" lemma="dann">dann</w>
  <w n="anschließend" pos="ADV" lemma="anschließend">anschließends</w>
  <pause duration="micro"/>
  <w pos="ADV" lemma="daraus">daraus</w>
  <w pos="VVFIN" lemma="folgen">folgt</w>
  <time time="2.98"/>
  <breathe type="in" length="1"/>
  <w n="äh" pos="SIEITJ" lemma="äh">öh</w>
  <time time="4.00"/>
</contribution>
```

Figure 2: Transcript XML excerpt (simplified)

All transcriptions are double checked by another student assistant and by the project coordinator before they are passed on to the next stage.

3.3 Annotation

Whereas the modified orthography used in transcription is necessary to adequately represent phenomena of speech in a manner that is suitable for (mostly qualitative) conversation analytic approaches to the data, it also makes further automatic processing and querying of the data more difficult, because the variety of forms in which a single lexical item occurs in the corpus becomes difficult to predict (consider, for example, the nine different forms *nein*, *nee*, *na*, *ne*, *neeh*, *nehee*, *nö*, *näh* and *nää* so far used to transcribe instances of the negation particle *nein*).
A second annotation layer is therefore added to the transcription in which each form is mapped onto the corresponding form (or forms, plural, in cases of contractions like "haste" → "hast Du") in standard orthography. This process is carried out semi-automatically, starting with a lookup in a lexicon of

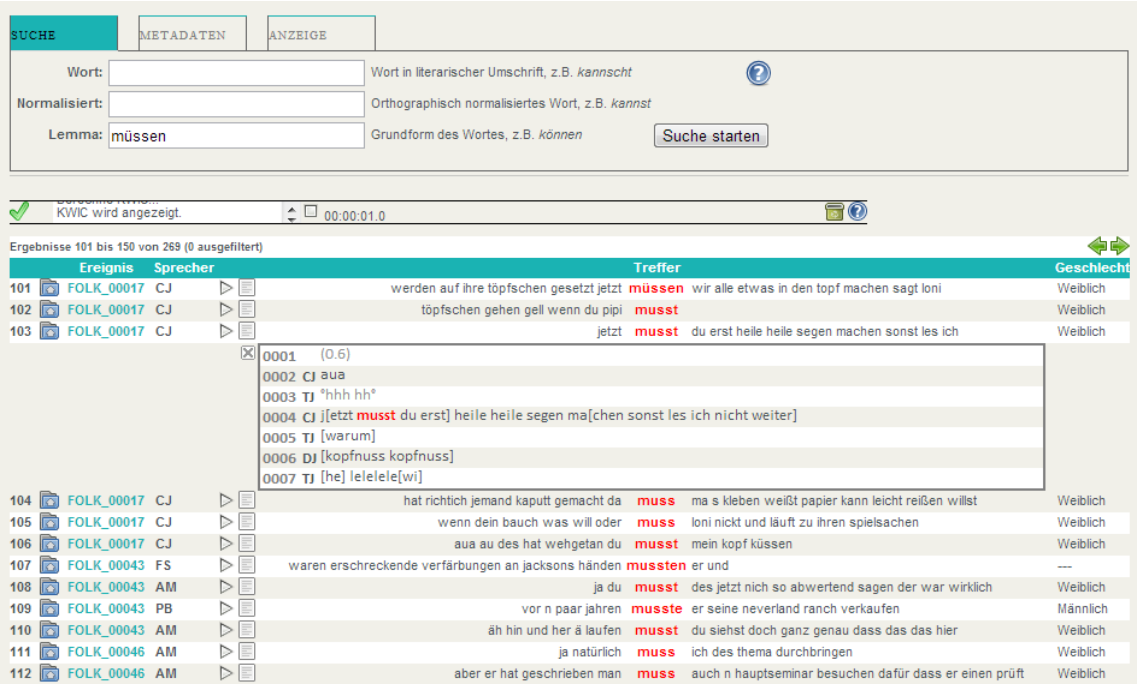

Figure 4: Query for lemma ‘müssen’ in the DGD2

previously annotated items, which yields around 80% correct mappings. The remaining 20% of erroneous normalisations are then corrected manually using the OrthoNormal tool (Schmidt 2012), which is optimised for this task.

A third and fourth layer of annotation are added by using TreeTagger (Schmid 1995) on the normalised version of the data. The resulting lemmatisation has low error rates (less than 2%) and is used without further manual correction. The POS tagging, in a first version based on the STTS tagset and a parameter file for written (newspaper) data, currently still produces error rates of around 20%, which is not an acceptable value for our purposes. We have therefore started working both on a modified version of the tagset which is better suited for spontaneous spoken language data (Westpfahl & Schmidt 2014, Zinsmeister et al. 2014) and on an adaptation of the TreeTagger parameter file for spoken data.

Figure 3 illustrates the different annotation layers of FOLK.

Transcription	da	gehst	de	jetz	einfach	über	dem	bild
Normalisation	da	gehst	Du	jetzt	einfach	über	dem	Bild
Lemmatisation	da	gehen	du	jetzt	einfach	über	d	Bild
POS	ADV	VFIN	PPER	ADV	ADJD	APPR	ART	NN

Figure 3: Annotation levels in FOLK

3.4 Metadata

Metadata about interactions and speakers are transferred from the paper forms gathered in the field to XML files following the metadata schema of the Archive of Spoken Language (Gasch 2008). Appropriate technical and archival metadata are added in this process. A data set is complete and ready for publication when all necessary consent forms, metadata documentations and transcription and annotation layers are available and have been checked for quality and consistency.

4. Dissemination

FOLK primarily addresses students and researchers in conversation analysis and related fields, as well as corpus linguists with an interest in spoken language. In terms of dissemination, this means that much attention has to be paid to facilitating access for groups of people without a strong background in computer science.

The FOLK corpus is published via the Database for Spoken German (DGD2, Schmidt 2014). The DGD2 includes functionality for browsing metadata, transcripts and recordings of the corpus, for systematically querying transcripts and metadata (see figure 4), and for downloading excerpts and selected full datasets. Usage of the DGD2 is free for non-commercial research and teaching purposes in academia. A one-time registration is required. The first version of the FOLK corpus was published with a beta release of the DGD2 in February 2012. The current version, published in March 2014, contains data for 137 interactions totalling altogether 101 hours of speech or 965,652 transcribed word tokens. As testified by more than 2000 registered users for the DGD2 so far, more than 60% of which are mainly working with FOLK, community interest in the corpus is great.

5. Outlook

Since development of the corpus is one of the institute’s permanent projects, FOLK will continue to grow over the coming years, both in terms of quantity of recordings and transcriptions and in terms of diversity of interaction

types.

Current work focuses on the acquisition of more data from the public space (e.g. panel discussions), on further balancing the corpus in terms of regional distribution, and on the integration of data from driving lessons.

As the project continues, the corpus construction workflow will be further optimized. Transcription still being the major bottleneck preventing FOLK from growing more quickly, we have started to experiment with speech technology (pause detection, speech recognition) to speed up the transcription process. So far, however, these experiments were successful only for a small range of data.

The dissemination methods will also be further improved. One of the next challenges in this respect will be the integration of video data into the corpus platform.

6. Acknowledgements

FOLK is developed by members of the Archive of Spoken German (AGD) at the Institute for the German Language (IDS) in Mannheim. We gratefully acknowledge data donations by the projects "Gesprochene Wissenschaftssprache Kontrastiv" (University of Leipzig, Christian Fandrych), "Sprachvariation in Norddeutschland" (University of Hamburg, Ingrid Schröder and colleagues from Universities Kiel, Münster, Potsdam, Frankfurt/O. and Bielefeld) and "Languages and Emotion" (Free University Berlin, Margret Selting) and by Kathrin Hee.

7. References

Berens, F.-J.; Jäger, K.-H.; Schank, G. and Schwitalla, J. (1976). *Projekt Dialogstrukturen. Ein Arbeitsbericht.* Heutiges Deutsch I/12. München: Hueber

Brinckmann, C.; Kleiner, S.; Knöbl, R. and Berend, N. (2008): *German Today: an areally extensive corpus of spoken Standard German.* In: Proceedings of LREC 2008, Marrakech, Morocco.

Engel, U. and Vogel, I. (1975) (Eds.). *Gesprochene Sprache.* Bericht der Forschungsstelle Freiburg. Tübingen: Narr.

Fandrych, C.; Meißner, C. and Slavcheva, A. (2012): The GeWiss Corpus: Comparing Spoken Academic German, English and Polish. In: T. Schmidt & K. Wörner (eds.): *Multilingual corpora and multilingual corpus analysis.* Amsterdam:Benjamins, pp. 319 – 337.

Gasch, Joachim (2008): *XML Schema driven Database Management of Speech Corpus Metadata.* In: SDV - Sprache und Datenverarbeitung/International Journal for Language Data Processing. Vol. 32.1/2008, pp. 23-33.

Hee, K. (2012): *Polizeivernehmungen von Migranten: Eine gesprächsanalytische Studie interkultureller Interaktionen in Institutionen.* Universitätsverlag Winter.

Kellner, B.; Lehmberg, T.; Schröder, I.; Wörner, K. (2008): *Data structures for the analysis of regional language variation.* In: A. Storrer; A. Geyken; A. Siebert & K.Würzner (eds.): *Text Resources and Lexical Knowledge.* Berlin: Mouton de Gruyter, pp. 53-63.

Schmid, H. (1995): *Improvements in Part-of-Speech Tagging with an Application to German.* Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Schmidt, T. (2012). *EXMARaLDA and the FOLK tools.* In: Proceedings of LREC 2012, Istanbul, Turkey.

Schmidt, T. (2014). *The Database for Spoken German – DGD2.* In: Proceedings of LREC 2014, Reykjavik, Iceland.

Selting, M. et al. (2009): *Gesprächsanalytisches Transkriptionssystem 2 (GAT 2).* In: Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion 10 (2009), pp. 353-402.

Westpfahl, S. & Schmidt, T. (2013): *POS für(s) FOLK– Part of Speech Tagging des Forschungs- und Lehrkorpus Gesprochenes Deutsch.* To appear in: Journal for Language Technology and Computational Linguistics.

Wiese, H. (2012): *Kiezdeutsch. Ein neuer Dialekt entsteht.* München 2012.

Zinsmeister, H; Heid, U. & Beck, K. (2014): *Adapting a part-of-speech tagset to non-standard text: the case of STTS.* In: Proceedings of LREC 2014, Reykjavik, Iceland.